Evaluating the Social Impact of Generative AI Systems in Systems and Society

Year: 2023 | *Citations: 143* | *Authors: Irene Solaiman, Zeerak Talat, William Agnew*

Abstract

Generative AI systems across modalities, ranging from text (including code), image, audio, and video, have broad social impacts, but there is no official standard for means of evaluating those impacts or for which impacts should be evaluated. In this paper, we present a guide that moves toward a standard approach in evaluating a base generative AI system for any modality in two overarching categories: what can be evaluated in a base system independent of context and what can be evaluated in a societal context. Importantly, this refers to base systems that have no predetermined application or deployment context, including a model itself, as well as system components, such as training data. Our framework for a base system defines seven categories of social impact: bias, stereotypes, and representational harms; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs. Suggested methods for evaluation apply to listed generative modalities and analyses of the limitations of existing evaluations serve as a starting point for necessary investment in future evaluations. We offer five overarching categories for what can be evaluated in a broader societal context, each with its own subcategories: trustworthiness and autonomy; inequality, marginalization, and violence; concentration of authority; labor and creativity; and ecosystem and environment. Each subcategory includes recommendations for mitigating harm.